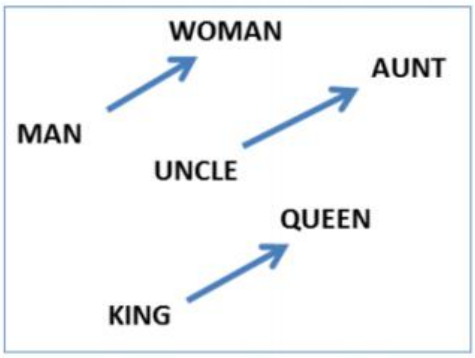
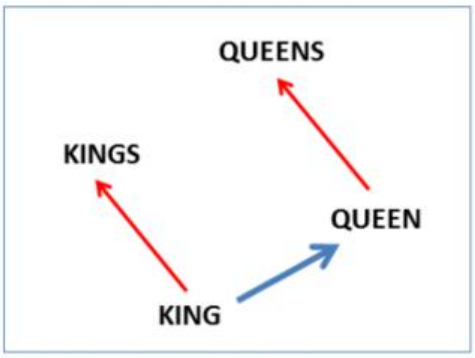
引用率过万的论文出错？从「词类比」说起

人工智能7 月 24 日

2013 年 ， Tomas Mikolov 发 表 的 《 Efficient estimation of word representations in vector space》，目前引用率已经超过 11K。除了其词向量的贡献外，一个让人印象深刻的贡献便是指出 NLP 中「词类比」的现象，最经典的例子莫过于「国王-男人+女人=皇后」。





Mikolov 在另外一篇引用率极高的文章《Linguistic regularities in continuous space word representations》中也着重强调了在连续空间词表示的语言规律。此后，关于词类比的研究此起彼伏，有诸多相关论文发表，雷锋网 AI 科技评论在一周前也曾发表过一篇 ACL 2019 论文解读，介绍词类比的理论解释：「国王-男人+女人=皇后」背后的词类比原理究竟为何？| ACL 2019。

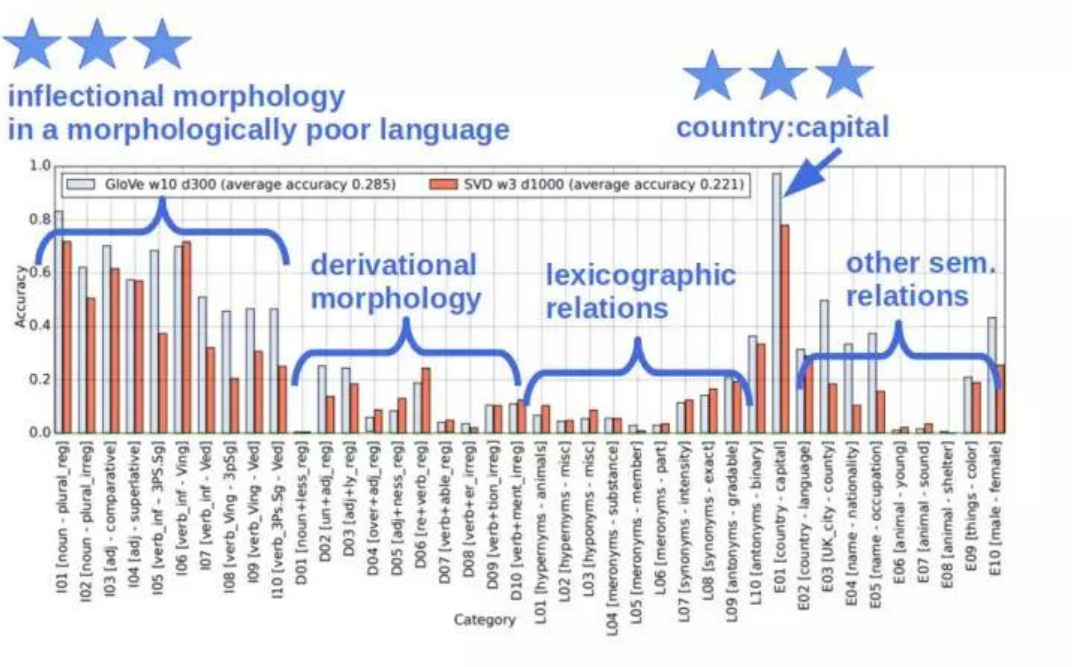
但是，在 NLP 中太过美好的事情往往都存在潜在的风险。

马萨诸塞大学（洛厄尔分校）文本机器实验室的 Anna Rogers 近日发表一篇博客，指出了词类比存在的问题以及由此引发的「如何让错误结论停止传播」的问题，值得我们思考。

1、词类比存在的问题

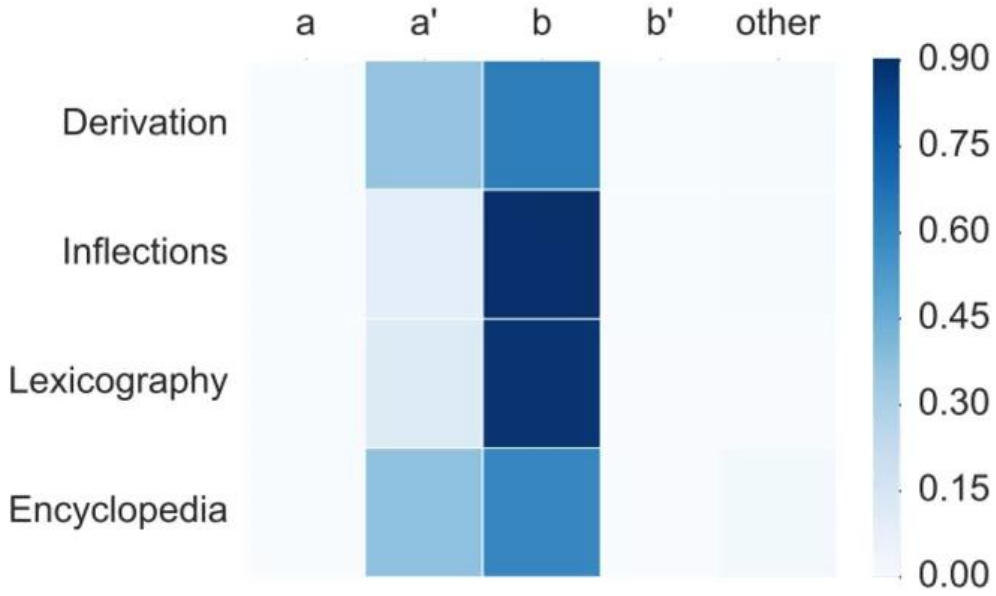
据我所知，首次对矢量偏移（vector offset）提出质疑的是 Köper 等人发现它在词典关系（ lexicographic relations ）中并不适用[1]，后来 Karpinska 等人证实了这个结论[2]。

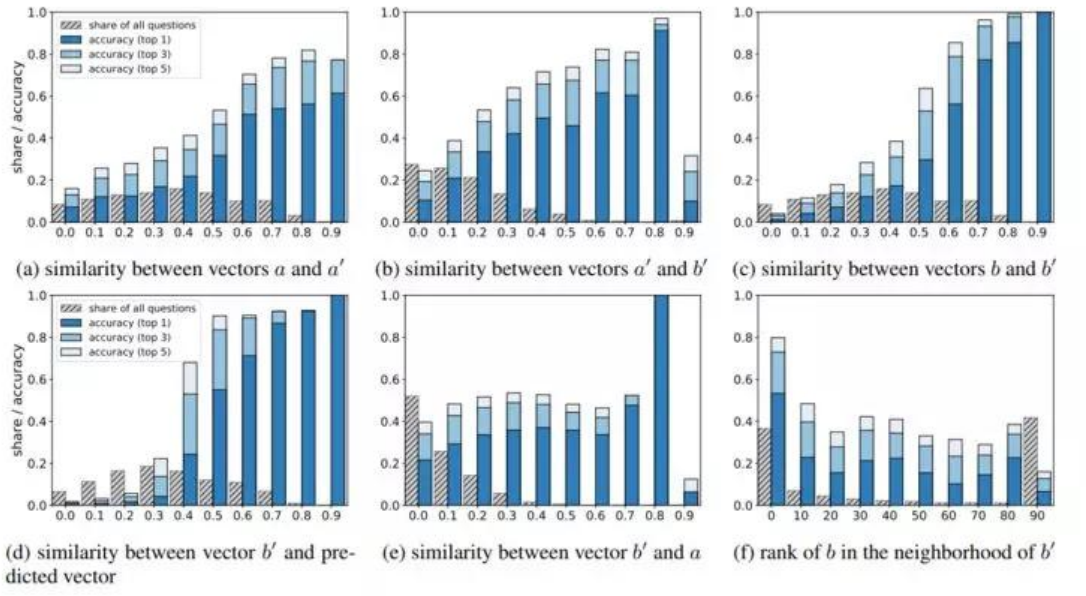
之后，Gladkova 等人的工作发现，BATS 数据集提供的包含 40 类关系的更大平衡样本中，矢量偏移仅适用于恰好包含原始 Google 数据集中的那部分[3]。如下图所示，40 类关系中仅「形态较差语言中的屈折形态」和「国家 ： 首都」这类词才能取得较高的准确率。



如果语言关系能够如 Mikolov 等人文章中所说的那样整齐和规律，那么为什么这个规律（词类比）不能概括所有词呢？

一些研究工作证明，如果 3 个源词（source words）没有从待选答案集中排除的话，词类比就不会起作用。例如在 king-man+woman=queen 的这个结论中，king、man 和 woman 的向量是被排除在可能的答案集之外的。Tal Linzen 的工作[4]表明，不用词类比，你只需要简单地获取 woman 最近邻的词，或者同时与 woman 和 king(没有 man)最相似的词，便可以获得相当高的准确率。在 Rogers 等人[5]的工作中指出，如果你不排除 3 个源词的话会发生什么：





(a) similarity between vectors a and a' (b) similarity between vectors a' and b' (c) similarity between vectors b and b'

(d) similarity between vector b' and predicted vector (e) similarity between vector b' and a (f) rank of b in the neighborhood of b'

其中 a、a’ 、b 是源向量，b’ 是目标向量。可以看出，在大多数情况下词类比的最好结果竟然是 b（也即 woman）。

如果在大多数情况下预测向量都是最接近 woman 向量，这就意味着矢量偏移太小，偏移本身并没有产生实质性意义，你的结果仍然停留在源向量附近。

Rogers 等人的研究还指出 ,如果源向量 a(“man”),a’(king)和 b(“woman”)被排除在外，那么你成功的可能性取决于正确答案与源词的接近程度，从下图可以看出：

你可以能会反对说：出现以上问题的原因是不良的词嵌入，理想的嵌入能够编码所有可能的关系以便能够通过矢量偏移来得到目标向量。

这种反对目前来看，只能期望通过未来的实验来验证了。

但从理论角度来看，即使理想的嵌入也不可能得出通用的词类比关系，原因如下：

（1）从语义角度，操纵向量差的想法让人想起上世纪 50 年代的成分分析方法，针对成分分析已经有充足的理由来说明为什么不值得继续发展，例如“man” + “unmarried”作为“bachelor”（单身汉）的定义是否适用于“Pope”（教皇）？

（2）从分布角度，即使看似完美的类比（如，cat:cats 与 table:tables）也并不完美。例如 turn the tables （翻桌子）与 turn the table（转过桌子）并不相似，它们出现在不同的上下文中，而这种差异在 cat:cats 中却不存在。鉴于这样的差异成千上万，我们怎么能够期望总体能够表现出完美的类比规则呢？如果真的这样做了，它们能够很好地代表语言语义吗？如果我们想获得良好的语言生成，我们就需要考虑到这种细微的差异，而不是粗暴地忽略它们。

总结来说，以上几篇论文对怀疑矢量偏移效果提供了充分的理由。矢量偏移似乎更适用于小的原始数据集，前提是预测目标要事先排除掉源向量；其成功的部分原因可归结为基本余弦相似性，但它无法概括为更广泛的语言关系。

2、欠缺的影响力

我写这篇文章的重点，想说的并不仅仅是上面提到的关于矢量偏移的负面证据，而是这些负面结果以及相关的报告从来没有被受 Mikolov 论文影响的那成千上万的研究者所广泛了解。

这种现象也很容易理解。对于一个广泛传播的谣言，即使后期有诸多辟谣，也无法覆盖所有被影响的人。因此，辟谣是重要的，对辟谣的广泛支持和传播更为重要。

在科学领域，如果对一篇被广泛引用但有瑕疵的论文的结论进行更新，那么快速传播这种更新的结论符合每个研究人员的利益，这可以节省更多研究人员浪费在原始未经测试的假设上的努力。

然而不幸的是，以上提到的那些研究成果，仅有一篇发表在顶会上（Schluter, NAACL 2018），这或许并非巧合。作为对比，现在已经有两篇 ACL 论文、一篇 COLING 论文和 ICML 的一篇最佳论文为矢量偏移能起作用提供数学证明[6][7][8][9]。注意，Schluter 的论文也是采用了数学的观点，却得出了完全相反的结论。

当然我对矢量偏移持完全开放的态度，它有可能是对的，但也可能是错的。如果前者，那么说明我们拥有了一个直观、方便且可靠的方法来进行类比推理。但必须要强调的是，目前那些证明矢量偏移有效的论文并没有解决它的负面证据。

考虑假如上面的那些负面证据是正确的，那对该领域该有多大的影响？这意味着我们大多数人正在追求一个简单却不真实的语言关系模型，许多从业者在实际工作中也在使用这种方法。

总结：类比推理是人类推理中一个非常重要的方面，如果我们要达到通用人工智能，我们必须做到正确。截止目前为止，从我所看到的，词嵌入的线性矢量偏移并不是正确的思考方式。但除了它，还有许多其他的方向，包括一些更好的推理方法[][]，或许我们也该尝试一下其他更有希望的方向。

3、如何让「谣言止于智者」

矢量偏移的问题并不是个别现象。它是一类模式的代表：（1）有一个闪亮的结果，直观、有吸引力，然后又因为过于出名而少有质疑；（2）负面的结果可见度低，并不为大多数人所注意。

在 NLP 领域，后者因为近年来 Arxiv 论文暴涨而加剧。当你连自己想要阅读的论文列表都读不完时，哪还有心思去关注哪些小众的引用率低的论文？最自然的选择就是，重点关注引用率最高的哪些。

事实上，很难让负面结果变得如那些明星论文一样性感，正如辟谣从来没有谣言传播力大一样。

但我认为，可以通过某种机制来改善这种情况。为什么我们不在 ACL 这样会议上设立负面结果的奖励呢，这可以鼓励人们对那些被广泛接受的假设进行事实核查？这将：

- 提高对流行问题的认识，使人们不会在不牢靠的假设基础上进行进一步工作;
- 确定明年需要更多人手的高产研究方向，从而刺激 NLP 的整体进展;
- 通过鼓励研究和报告负面结果来减少错误重复的问题。

例如 NAACL 2019 上就有几篇有意思的论文就可以获得此种类型的奖：

- exposing the lack of transfer between QA datasets (Yatskar, 2019)

limitations of attention as “explaining” mechanism (Jain & Wallace, 2019)

-
- multimodal QA systems that work better by simply ignoring some of the input modalities (Thomason, Gordon, & Bisk, 2019)

-

这三篇论文中有两篇都只是 poster paper。我无法想象有多少类似的工作甚至都没有通过评审。我觉得这对做类似重要工作的人发出了一个错误的信号，告诉他们明年不要做这种类型的工作了。很悲哀！

想象一下，假如有这样一个奖，并且被授予给 Yatskar。那么参加这个会议的每个人（甚至更多人）都会知道三个流行的问答数据集之间缺乏迁移。QA 是最流行的任务之一，所有如果能够让整个社区知道这个问题，来年就会有更多的人去解决 QA 中的这个问题，而不是单纯地集中在某一个数据集上进行研究。

负面结果的论文，应当被重视，也应当被强调！

参考资料：

[1] Köper, M., Scheible, C., & im Walde, S. S. (2015). Multilingual Reliability and "Semantic" Structure of Continuous Word Spaces. Proceedings of the 11th International Conference on Computational Semantics, 40 – 45. Association for Computational Linguistics.

[2] Karpinska, M., Li, B., Rogers, A., & Drozd, A. (2018). Subcharacter Information in Japanese Embeddings: When Is It Worth It? Proceedings of the

Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP, 28–37. Melbourne, Australia: Association for Computational Linguistics.

[3] Gladkova, A., Drozd, A., & Matsuoka, S. (2016). Analogy-Based Detection of Morphological and Semantic Relations with Word Embeddings: What Works and What Doesn't. Proceedings of the NAACL-HLT SRW, 47–54. https://doi.org/10.18653/v1/N16-2002

[4] Linzen, T. (2016). Issues in Evaluating Semantic Spaces Using Word Analogies. Proceedings of the First Workshop on Evaluating Vector Space Representations for NLP. https://doi.org/http://dx.doi.org/10.18653/v1/W16-2503

[5]Rogers, A., Drozd, A., & Li, B. (2017). The (Too Many) Problems of Analogical Reasoning with Word Vectors. Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017), 135–148.

[6] Gittens, A., Achlioptas, D., & Mahoney, M. W. (2017). Skip-Gram - Zipf + Uniform = Vector Additivity. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 69–76. https://doi.org/10.18653/v1/P17-1007

[7] Hakami, H., Hayashi, K., & Bollegala, D. (2018). Why Does PairDiff Work? - A Mathematical Analysis of Bilinear Relational Compositional Operators for Analogy Detection. Proceedings of the 27th International Conference on Computational Linguistics, 2493–2504.

[8] Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). Towards Understanding Linear Word Analogies. To Appear in ACL 2019.

[9] Allen, C., & Hospedales, T. (2019). Analogies Explained: Towards Understanding Word Embeddings. ArXiv:1901.09813 [Cs, Stat].

- END -